

Google DeepMind's AlphaMissense classifies 71 million possible disease-linked mutations

Source copy kept by wheresthe.ai for the story "Google DeepMind's AlphaMissense classifies 71 million possible disease-linked mutations" (<https://wheresthe.ai/d/deepmind-alphamissense-2023-09/>).

Source: <https://deepmind.google/discover/blog/a-catalogue-of-genetic-mutations-to-help-pinpoint-the-cause-of-diseases/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 6d18260865e30360c65676d53d657dcd35a6e30ada3c6470008201452c86e6f6

September 19, 2023 Science A catalogue of genetic mutations to help pinpoint the cause of diseases

Share

New AI tool classifies the effects of 71 million 'missense' mutations Uncovering the root causes of disease is one of the greatest challenges in human genetics. With millions of possible mutations and limited experimental data, it's largely still a mystery which ones could give rise to disease. This knowledge is crucial to faster diagnosis and developing life-saving treatments. Today, we're releasing a catalogue of 'missense' mutations where researchers can learn more about what effect they may have. Missense variants are genetic mutations that can affect the function of human proteins. In some cases, they can lead to diseases such as cystic fibrosis, sickle-cell anaemia, or cancer. The AlphaMissense catalogue was developed using AlphaMissense, our new AI model which classifies missense variants. In a paper published in Science, we show it categorised 89% of all 71 million possible missense variants as either likely pathogenic or likely benign. By contrast, only 0.1% have been confirmed by human experts. AI tools that can accurately predict the effect of variants have the power to accelerate research across fields from molecular biology to clinical and statistical genetics. Experiments to uncover disease-causing mutations are expensive and laborious – every protein is unique and each experiment has to be designed separately which can take months. By using AI predictions, researchers can get a preview of results for thousands of proteins at a time, which can help to prioritise resources and accelerate more complex studies. We've made all of our predictions freely available for commercial and researcher use, and open sourced the model code for AlphaMissense.

AlphaMissense predicted the pathogenicity of all possible 71 million missense variants. It classified 89% – predicting 57% were likely benign and 32% were likely pathogenic.

What is a missense variant? A missense variant is a single letter substitution in DNA that results in a different amino acid within a protein. If you think of DNA as a language, switching one letter can change a word and alter the meaning of a sentence altogether. In this case, a substitution changes which amino acid is translated, which can affect the function of a protein. The average person is carrying more than 9,000 missense variants. Most are benign and have little to no effect, but others are pathogenic and can severely disrupt protein function. Missense variants can be used in the diagnosis of rare genetic diseases, where a few or even a single missense variant may directly cause disease. They are also important for studying complex diseases, like type 2 diabetes, which can be caused by a combination of many different types of genetic changes. Classifying missense variants is an important step in understanding which of these protein changes could give rise to disease. Of more than 4 million missense variants that have been seen already in humans, only 2% have been annotated as pathogenic or benign by experts, roughly 0.1% of all 71 million possible missense variants. The rest are considered 'variants of unknown significance' due to a lack of experimental or clinical data on their impact. With AlphaMissense we now have the clearest picture to date by classifying 89% of variants using a threshold that yielded 90% precision on a database of known disease variants. Pathogenic or benign: How AlphaMissense classifies variants

AlphaMissense is based on our breakthrough model AlphaFold, which predicted structures for nearly all proteins known to science from their amino acid sequences. Our adapted model can predict the pathogenicity of missense variants altering individual amino acids of proteins. To train AlphaMissense, we fine-tuned AlphaFold on labels distinguishing variants seen in human and closely related primate populations. Variants commonly seen are treated as benign, and variants never seen are treated as pathogenic. AlphaMissense does not predict the change in protein structure upon mutation or other effects on protein stability. Instead, it leverages databases of related protein sequences and structural context of variants to produce a score between 0 and 1 approximately rating the likelihood of a variant being pathogenic. The continuous score allows users to choose a threshold for classifying variants as pathogenic or benign that matches their accuracy requirements.

An illustration of how AlphaMissense classifies human missense variants. A missense variant is input, and the AI system scores it as pathogenic or likely benign. AlphaMissense combines structural context and protein language modelling, and is fine-tuned on human and primate variant population frequency databases.

AlphaMissense achieves state-of-the-art predictions across a wide range of genetic and experimental benchmarks, all without explicitly training on such data. Our tool outperformed other computational methods when used to classify variants from ClinVar, a public archive of data on the relationship between human variants and disease. Our model was also the most accurate method for predicting results from the lab, which shows it is consistent with different ways of measuring pathogenicity.

AlphaMissense outperforms other computational methods on predicting missense variant effects. Left: Comparing AlphaMissense and other methods' performance on classifying variants from the Clinvar public archive. Methods shown in grey were trained directly on ClinVar and their performance on this benchmark are likely overestimated since some of their training variants are contained in this test set. Right: Graph comparing AlphaMissense and other methods' performance on predicting measurements from biological experiments.

Building a community resource AlphaMissense builds on AlphaFold to further the world's understanding of proteins. One year ago, we released 200 million protein structures predicted using AlphaFold – which is helping millions of scientists around the world to accelerate research and pave the way toward new discoveries. We look forward to seeing how AlphaMissense can help solve open questions at the heart of genomics and across biological science. We've made AlphaMissense's predictions freely available to both commercial and scientific communities. Together with EMBL-EBI, we are also making them more usable through the Ensembl Variant Effect Predictor . In addition to our look-up table of missense mutations, we've shared the expanded predictions of all possible 216 million single amino acid sequence substitutions across more than 19,000 human proteins. We've also included the average prediction for each gene, which is similar to measuring a gene's evolutionary constraint – this indicates how essential the gene is for the organism's survival.

Examples of AlphaMissense predictions overlaid on AlphaFold predicted structures (red=predicted as pathogenic, blue=predicted as benign, grey=uncertain). Red dots represent known pathogenic missense variants, blue dots represent known benign variants from the ClinVar database. Left: HBB protein. Variants in this protein can cause sickle cell anaemia. Right: CFTR protein. Variants in this protein can cause cystic fibrosis.

Accelerating research into genetic diseases A key step in translating this research is collaborating with the scientific community. We have been working in partnership with Genomics England, to explore how these predictions could help study the genetics of rare diseases. Genomics England cross-referenced AlphaMissense's findings with variant pathogenicity data previously aggregated with human participants. Their evaluation confirmed our predictions are accurate and consistent, providing another real-world benchmark for AlphaMissense. While our predictions are not designed to be used in the clinic directly – and should be interpreted with other sources of evidence – this work has the potential to improve the diagnosis of rare genetic disorders, and help discover new disease-causing genes. Ultimately, we hope that AlphaMissense, together with other tools, will allow researchers to better understand diseases and develop new life-saving treatments.

Learn more about AlphaMissense:

Read our paper in Science Ensembl Variant Effect Predictor plugin Download AlphaMissense code

Notes *As of 13 March 2024 the AlphaMissense predictions are available under a CC BY v.4 license, thereby lifting the previous non-commercial use restriction. Please see published database and Zenodo for further access information. We would like to thank Juanita Bawagan, Jess Valdez, Katie McAtackney, Kathryn Seager, Hollie Dobson, for their help with text and figures. We are also grateful to our external partners, Genomics England and EMBL-EBI, for their continuous support. This work was done thanks to the contributions of the co-authors: Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli. We would also like to thank Kathryn Tunyasuvunakool, Rob Fergus, Eliseo Papa, David La, Zachary Wu, Sara-Jane Dunn, Kyle R. Taylor, Natasha Latysheva, Hamish Tomlinson, Augustin Žídek, Roz Onions, Mira Lutfi, Jon Small, Molly Beck, Annette Obika, Hannah Gladman, Folake Abu, Alyssa Pierce, James Tam, Q Green, Meera Last, Tharindi Hapuarachchi and the greater Google DeepMind team for their support, help and feedback.